Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting

Year: 2023 | Citations: 518 | Authors: Melanie Sclar, Yejin Choi, Yulia Tsvetkov, Alane Suhr

---

Abstract

As large language models (LLMs) are adopted as a fundamental component of language technologies, it is crucial to accurately characterize their performance. Because choices in prompt design can strongly influence model behavior, this design process is critical in effectively using any modern pre-trained generative language model. In this work, we focus on LLM sensitivity to a quintessential class of meaning-preserving design choices: prompt formatting. We find that several widely used open-source LLMs are extremely sensitive to subtle changes in prompt formatting in few-shot settings, with performance differences of up to 76 accuracy points when evaluated using LLaMA-2-13B. Sensitivity remains even when increasing model size, the number of few-shot examples, or performing instruction tuning. Our analysis suggests that work evaluating LLMs with prompting-based methods would benefit from reporting a range of performance across plausible prompt formats, instead of the currently-standard practice of reporting performance on a single format. We also show that format performance only weakly correlates between models, which puts into question the methodological validity of comparing models with an arbitrarily chosen, fixed prompt format. To facilitate systematic analysis we propose FormatSpread, an algorithm that rapidly evaluates a sampled set of plausible prompt formats for a given task, and reports the interval of expected performance without accessing model weights. Furthermore, we present a suite of analyses that characterize the nature of this sensitivity, including exploring the influence of particular atomic perturbations and the internal representation of particular formats.